**Got Data? Use Spark Streaming in Azure Databricks**

Challenge 3 – An Introduction Sparking Streaming

What You’ll Need

To complete the labs, you will need the following:

• A web browser

• A Microsoft account

• A Microsoft Azure subscription

• A Windows, Linux, or Mac OS X computer

• Azure Storage Explorer

• The lab files for this course <only if you get stuck – proctors have access to this content>

**Note**: To set up the required environment for the lab, follow the instructions in the [**Setup**](https://microsoft-my.sharepoint.com/personal/laedell_microsoft_com/Documents/Hackathons/ML/Databricks/databricks_intro/databricks-introFinal/Setup%20Guide.docx)document for

this course. Specifically, you must have signed up for an Azure subscription.

Challenge Background:

Spark structured streaming enables you to use the dataframe API to read and process an unbounded

stream of data. This kind of processing is used in real-time scenarios to aggregate data over temporal

intervals or *windows*. You can use Spark to process streaming data from a wide range of sources,

including Azure Event Hubs, Kafka, and others. In this lab, you’ll process data as it is added to a folder in

Azure blob storage.

Upload Initial Data to Azure Storage

In this exercise, you will process a stream of data that simulates status information generated by

Internet-of-things (IoT) devices. The data will be written to a blob storage container where it can be

accessed by your Spark cluster. The instructions here assume you will use Azure Storage Explorer to

upload the data, but you can use any Azure Storage tool you prefer.

Business Challenge: Find business value from large stream of data that simulates status information generated by

Internet-of-things (IoT) devices

Questions:

1. What types of analyses would be good when working with website log data ? Remember, you want to provide value to the business quickly.
2. If you look at the ProcessLog.py files, what Spark modules are being used and why do you think they were chosen? Would you have chosen alternate modules and why?
3. What else would you add to the script to enhance what it is already doing using what you learned in Challenge 1?
4. Can you explain the purpose of the different files that were created by the Spark job process in your blob storage?
5. What is the file name(s) created that contains the Processed Results? Does it contain the # of webpage views for each product that you were expecting?

**Challenge 3:**

3.1) Access Databricks Workspace from Challenge 1 or Provision a New Databricks Workspace if you didn’t complete Challenge 1

3.2) Access existing Storage Account & Create a new Container for Challenge 2 (if you didn’t complete Challenge 1, create a new Storage Account / Container)

**helpful tips:**

\*\*Upload the log source files - These logs have been made available as part of the Teams site under Files <IISLog.txt>

3.3) In the folder where you extracted the lab files for this course on your local computer, verify that

the **data\stream** folder contains four files named **stream\_*N*.txt**. These files contain simulated

status data from hypothetical IoT devices. Upload only **stream\_1.txt** as a block blob

to a new folder named **stream** in root of the **spark** container.

3.4) Process the Data Stream using Spark Streaming

You will use your choice of a Python or Scala script to process the streaming IoT data.

Source files containing the necessary code to process the data have been provided in Teams under Files.

Access Streaming.py or Streaming.scala – steps TBD

**helpful tips:**

TBD

**Challenge considerations:**

TBD

END OF CHALLENGE 3